PB85-153567

Computer Use in Research for Data Analysis

(U.S.) Environmental Protection Agency
Cincinnati, OH

Dec 84

U.S. Department of Commerce
National Technical Information Service
NTIS

PB85-153567

EPA-600/D-84-286
December 1984

COMPUTER USE IN RESEARCH FOR DATA ANALYSIS

by

Ben W. Lykins, Jr.

Susan Campbell

John C. Ireland

WATER ENGINEERING RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
CINCINNATI, OH 45268

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA *(Please read Instructions on the reverse before completing)*		
1. REPORT NO. EPA-600/D-84-286	2.	3. RECIPIENT'S ACCESSION NO. PB85 153567/AS
4. TITLE AND SUBTITLE COMPUTER USE IN RESEARCH FOR DATA ANALYSIS		5. REPORT DATE December 1984 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Ben W. Lykins, Jr., Susan Campbell, and John C. Ireland		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Water Engineering Research Laboratory-Cin., OH Office of Research and Development U.S. Environmental Protection Agency Cincinnati, OH 45268		10. PROGRAM ELEMENT NO. BNC1A 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Water Engineering Research Laboratory-Cin., OH Office of Research and Development U.S. Environmental Protection Agency Cincinnati, OH 45268		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

Data reported in various ways from several large field-scale research projects have been successfully incorporated into a structured format. This format is now required for all new projects that produce data in a machine-readable form. The database structure is general enough in design to accommodate a wide variety of project data scenarios. Remote entry of data from a research site (Thornton, Colorado) directly into the mainframe at NCC has been in progress for several months.

The data handling system is user-friendly and versatile. Data reduction on hundreds of thousands of records can be accomplished efficiently. These records can be analyzed, summarized, and reported in a matter of hours. Canned programs and command procedures allow the user to create, examine, update, modify, subset, superset, and plot or print all or a portion of the records contained within the main database. Modifications of the canned programs to accommodate special or unique situations can be accomplished with little effort.

17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS *(This Report)* Unclassified 20. SECURITY CLASS *(This page)* Unclassified	21. NO. OF PAGES 17 22. PRICE

EPA Form 2220-1 (9-73)

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

COMPUTER USE IN RESEARCH FOR DATA ANALYSIS

Ben W. Lykins, Jr.
Environmental Engineer

Susan Campbell
Statistical Assistant

John C. Ireland
Chemist

Drinking Water Research Division
Municipal Environmental Research Laboratory
U. S. Environmental Protection Agency
Cincinnati, OH 45268

Introduction

The Municipal Environmental Research Laboratory (MERL) of the U. S. Environmental Protection Agency (EPA) in Cincinnati, Ohio has been involved in various research activities to improve the quality of drinking water. Since 1976, the Drinking Water Research Division (DWRD) of MERL has conducted extramural field-scale research evaluating granular activated carbon (GAC) for the removal of trace organics from drinking water. This research effort produced thousands of pieces of data in various formats that were usable for little more than producing final project reports.

As the need arose to perform more or different evaluations of the data within a project or to compare projects, it became evident that a system had to be developed for making the data readily available and interactively manageable. To accomplish this required two major efforts: (1) development of a consistent data format, and (2) development of a database management and data retrieval and reporting system.

Data Collected at Project Sites

At each research project site, various amounts of data were collected. These data varied from routine, easily measureable parameters such as pH to specific organic compounds requiring sophisticated instrumentation for analysis. However, certain compounds and parameters were consistently measured in each project. Examples of these are basic operational parameters such as pH, temperature, chlorine residual, standard plate count and total coliforms and organics such as total organic carbon (TOC) and the trihalomethanes. Several of the project sites had various contaminants that were specific to that location. An example was Jefferson Parish, Louisiana, where atrazine and 1,2-dichloroethane were consistently detected.

Once a sampling and analysis format was developed for a project, data were collected over a long period (up to 2 years). During this time, various phases were established in which different treatment options were evaluated. Throughout this process many variables had to be considered in establishing a usable database management system.

Development of Consistent Data Format

Predictably, data were not reported in the same format for each research site. Reasons were that different project officers were responsible for the research and the utilities wanted to report data in a form useful to their particular operation. One example of a site-specific format came from Manchester, New Hampshire where one parameter (TOC) was reported for several locations in the treatment plant over time. At Passaic, New Jersey, the same parameter with its constituent parts was provided along with the percent removal for one date at several locations. Also, TOC was reported in mg/L for Manchester, and in ug/L for Passaic, (Table 1). At Evansville, Indiana several compounds were reported for one date and location. The same was true for the Ohio River Valley Sanitation Commission (ORSANCO) project (Huntington, West Virginia and Beaver Falls, Pennsylvania) with the addition of other information (Table 2).

Accumulated project data had to be computerized to handle the large volumes being reported. The first requirement was to develop a structured data format. Figure 1 shows the standardized coding form developed. These coding forms were consistent in structural content with unique variables in each field, except for the parmfull (expanded parameter name or designation) and note field. Before this coding form could be used, however, a common designation was required for each parameter (dictionary). For instance, chloroform was reported in various ways, but was given a consistent eight-character designation (Table 3). The dictionary also provided a description of all variables for a given field within each database.

As of December 1984 there are over 500,000 records containing the information shown in Figure 1 from 11 research projects in the DWRD database system. Table 4 shows the geographic locations and number of records for each research project.

Database Management

The DWRD/CIS (Drinking Water Research Division/Computer Information System) databases and associated programs reside at the EPA National Computer Center (NCC) in Research Triangle Park, North Carolina. The NCC is a centralized, large-scale (mainframe) computer operation that serves EPA users in Washington, D.C. headquarters, regional offices, and laboratories throughout the United States. The NCC primarily uses IBM equipment with one IBM 3081D and two IBM 370/168s coupled to an Amdahl V/6. TYMNET provides NCC with international and national access for interactive users. This access is gained in more than 400 U.S. cities through locally dialed numbers.

Users interact with the system through four major software components of the NCC installation: 1) the OS/MVS/XA (Operating System/Multiple Virtual Storage/Extended Architectural) operating system with OS/JCL (Operating System/Job Control Language) used to invoke programs and allocate disk and tape files, 2) the IBM Job Entry Subsystem (JES2) through which the bulk of the printed output of user-submitted jobs is produced, 3) Online Business System's WYLBUR text editor, which is used online to enter and edit raw data and construct programs by means of user terminals, and 4) IBM's Time Sharing Option (TSO), which allows users to access DWRD/CIS databases interactively through user terminals.

Data Access and Handling

As mentioned earlier, the typical DWRD/CIS user has access to and interacts with the NCC through dial-up remote terminals. DWRD currently uses several commercially popular microcomputers for this purpose, including Apples (all models), IBM PCs, a TRS-80 Model II, Texas Instruments Silent 700s, Anderson Jacobson 832's and several Tektronix 40xx series terminals, all of which are equipped with 300 and/or 1200 baud asynchronous modems. The Tektronix and Apple systems are interfaced with high resolution digital plotters to produce publication quality data plots, graphics, and text slides that may be generated with the available SASGraph or TELAGRAF mainframe graphics packages.

The workhorse of the DWRD/CIS is the Statistical Analysis System (SAS), a popular mainframe statistical package that resides on the NCC system. SAS is used sequentially to process the data from its raw (but machine-readable) form into SAS-format libraries from which it may be numerically analyzed, formatted, displayed, printed, or plotted. Canned SAS programs have been developed to interactively execute the most commonly required data reduction and analysis procedures. These canned procedures are invoked either through TSO command files, which ultimately drive the entire system, or through batch jobs, which may be submitted through the WYLBUR subsystem.

Required Outputs

During development of the computerized database, two major requirements were identified: (1) the ability to manipulate data within a research project, and (2) the ability to compare data between projects. Development of user-friendly canned programs and command procedures accomplished these objectives.

Outputs required from the database management system varied from simple percent removals, minimum, maximum, and average values to calculated values such as organic loading on carbon. Data values were required along with plots of data over time. In addition, the system had to be able to arrange data in columns for tabular presentation.

One example of data manipulation within a project appears in Figure 2 where the data show graphically the concentration of TOC collected at four points within a treatment system over a period of time. Another TOC example is presented in Figure 3. To produce the displayed loading curve for TOC, the canned SAS program needed input variables of TOC influent and effluent concentrations, the amount of water passed through the carbon, and the weight of carbon used in the study.

Comparison of data between projects seemed to be the major data handling challenge. Though all of the performance data could be identified by a consistent designation, sampling locations maintained their unique identity. Thus, knowledge was needed about the inherent specifics of the various projects in order to produce a comparative table or plot. This limitation in the system has not been completely eliminated, but it has been partially overcome by using note codes. The variability of the sampling locations was caused by failure to design the projects for comparative compatibility. However, in spite of these handicaps, data comparison between projects can be readily

accomplished. An example is presented in Figure 4, where five projects were compared for percent TOC removal over time.

An Optional Output

Various mathematical models have been developed for predicting GAC adsorptive capacity. Such models are frequently used to simulate batch adsorption experiments for single and bisolute systems. Typically, certain equilibrium expressions and constants are either estimated or calculated for a particular adsorbate-GAC combination. The equilibrium adsorption isotherms so determined are often extrapolated to predict full-scale GAC performance. Recently, an alternative strategy has been applied to model adsorption capacity characteristics by a non-linear least squares regression procedure. An easily customized batch routine was used to produce the regression and curve fitting shown in Figure 5.

Summary

Data reported in various ways from several large field-scale research projects have been successfully incorporated into a structured format. This format is now required for all new projects that produce data in a machine-readable form. The database structure is general enough in design to accommodate a wide variety of project data scenarios. Remote entry of data from a research site (Thornton, Colorado) directly into the mainframe at NCC has been in progress for several months.

The data handling system is user-friendly and versatile. Data reduction on hundreds of thousands of records can be accomplished efficiently. These records can be analyzed, summarized, and reported in a matter of hours. Canned programs and command procedures allow the user to create, examine, update, modify, subset, superset, and plot or print all or a portion of the records contained within the main database. Modifications of the canned programs to accommodate special or unique situations can be accomplished with little effort.

A major effort was required to transpose manually recorded data from completed and initiated research projects into a usable, consistent, and machine-readable format. The effort has been worthwhile, as evidenced by numerous requests and demands for the data. These data are readily accessible and have been used by the EPA Office of Drinking Water, Region VI, Drinking Water Research Division for producing summary reports on the use of GAC and biological activated carbon. The database has also been used for producing several other smaller reports and by contractual recipients of the American Water Works Research Foundation working on projects entitled: "Surrogate Parameters as an Indicator of Water Quality" and "Surface Water Quality Screening Model: A Case Study for Water Utility Management."

REFERENCES

1. Lykins, Jr., B.W., et al., "Granular Activated Carbon for Removing Nontrihalomethane Organics from Drinking Water," U.S. EPA, Cincinnati, Ohio (September 1984).

2. Kilpatrick, A.W., "DWRD Computer Information System User's Guide," Dynalectron Corporation Contract with U.S. EPA, Cincinnati, OH (December 1982).

3. Randtke, S.J. and Snoeyink, V.L., "Evaluating GAC Adsorptive Capacity," JAWWA, 75, 406-413 (August 1983).

4. Thacker, W.E., Snoeyink, V.L., and Crittenden, J.C., "Modeling of Activated Carbon and Coal Gasification Char Adsorbents in Single Solute and Bisolute Systems," Res. Rept. 161, Wtr. Resources Ctr., Univ. of Illinois, Urbana (July 1981).

5. Lee, M.C., Snoeyink, V.L. and Crittenden, J.C., "Activated Carbon Adsorption of Humic Substances," JAWWA, 73, 440 (August 6, 1981).

6. Crittenden, J.C., and Weber, W.J., Jr., "Predictive Model for Design of Fixed-Bed Adsorbers: Parameter Estimation and Model Development," Jour. Envir. Engrg. Div. ASCE, 104, 185 (April 1978).

7. Crittenden, J.C. and Weber, W.J., Jr., "Predictive Model for Design of Fixed-Bed Adsorbers: Single Component Model Verification," Jour. Envir. Engrg. Div. ASCE, 104, 443 (June 1978).

8. Crittenden, J.C., et al., "Mathematical Model of Sequential Loading in Fixed-Bed Adsorbers," Jour. WPCF, 52, 2780 (November 1980).

9. SAS Institute Inc., SAS User's Guide: Statistics, 1982 Edition, Cary, NC: SAS Institute Inc., pp. 15-37, 1982.

TABLE 1. EXAMPLES OF DATA REPORTING FORMATS FOR TOC

MANCHESTER

RUN: R32V3R

PARAMETER: TOC

DATE	DAY	RAW	PREMIX	POST CS	CARB. INF
2/16/81	0			3.1	2.8
2/17/81	1				2.6
2/19/81	3	4.6	4.1	2.9	2.5

PASSAIC

RUN1

DATE	SAMPLE SOURCE	ORGANIC CARBON TOC	VOC	NVOC	%REMOVAL
3/22/78	INF(DW)	1751	5	1746	
	F-1	590	1	589	66
	F-2	584	1	583	67
	F-3	669	1	668	62

TABLE 2. EXAMPLES OF DATA REPORTING FORMATS FOR ORGANIC COMPOUNDS

EVANSVILLE

I.D. JY780614-T3P-08737-MR
DATE 6/19/78

NO.	COMPOUND	UG/L.	NO.	COMPOUND	UG/L
23V	CHLOROFORM	22	48V	BROMO-DICHLORO-METHANE	7.7
6V	CARBON TET	<0.1	47V	BROMOFORM	0.3

ORSANCO

DATE: MAY 22, 1978 LOCATION: FINISHED

COMPOUND	STANDARD RETENTION TIME	SAMPLE RETENTION TIME	CONCENTRATION
CHLOROFORM	8.09	8.08	0.4
BROMO-DICHLOROMETHANE	11.34	11.32	0.4
CHLOROBENZENE	20.46	20.46	0.6

TABLE 3. PARAMETER ABBREVIATIONS FOR CHLOROFORM

LOCATION	PARM *	PARMFULL +
JEFFERSON PARISH	ICHC	CHLROFRM
PASSAIC	ICL3	CHLROFRM
MANCHESTER	CL3	CHLROFRM
MIAMI	ICL3	CHLROFRM
EVANSVILLE	23V	CHLROFRM
CINCINNATI	ICLR	CHLROFRM

* Abbreviated parameter as reported.

+ Expanded consistent parameter designation.

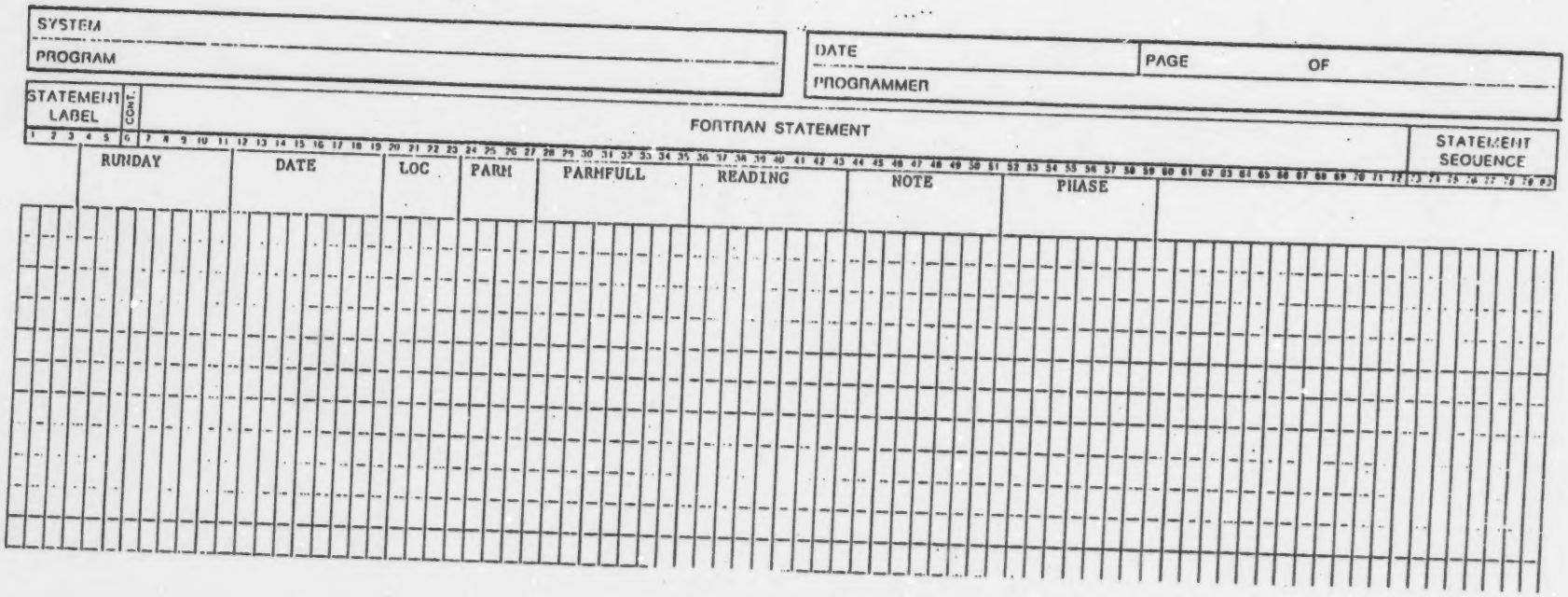

FIGURE 1. STANDARDIZED CODING FORM

TABLE 4. SIZE OF DRINKING WATER RESEARCH DIVISION DATABASE

LOCATION	NO. OF RECORDS
CINCINNATI, OH	94,572
JEFFERSON PARISH, LA (1)	243,463
PASSAIC VALLEY, NJ	7,046
EVANSVILLE, IN	50,583
MANCHESTER, NH	22,167
MIAMI, FL	19,565
ORSANCO (HUNTINGTON, BEAVER FALLS)	15,829
THORNTON, CO	17,765
SHREVEPORT, LA	13,348
PHILADELPHIA, PA	35,854
TOTAL	520,192

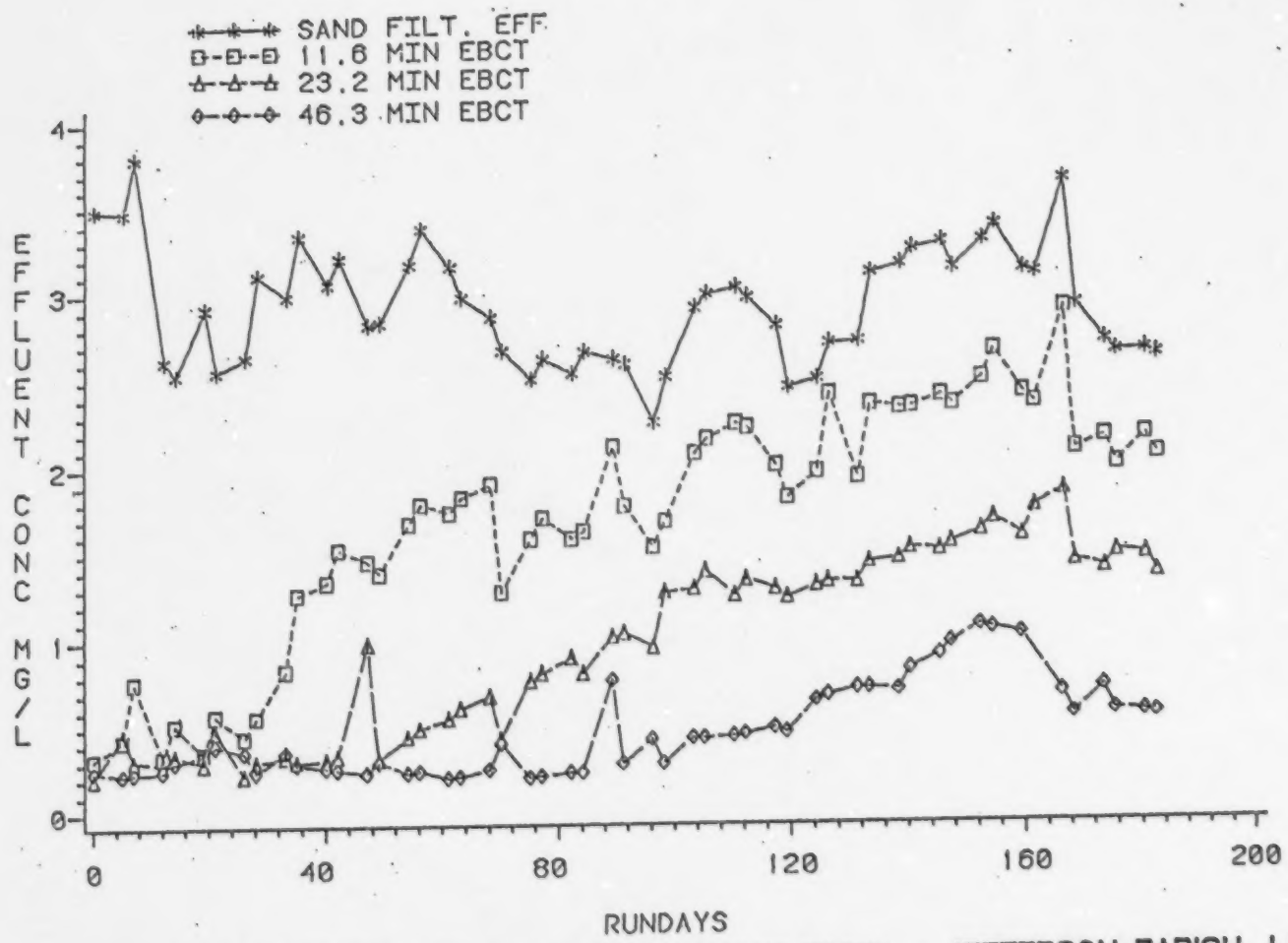

FIGURE 2. TOC REMOVAL FOR VARIOUS CONTACT TIMES – JEFFERSON PARISH, LA.

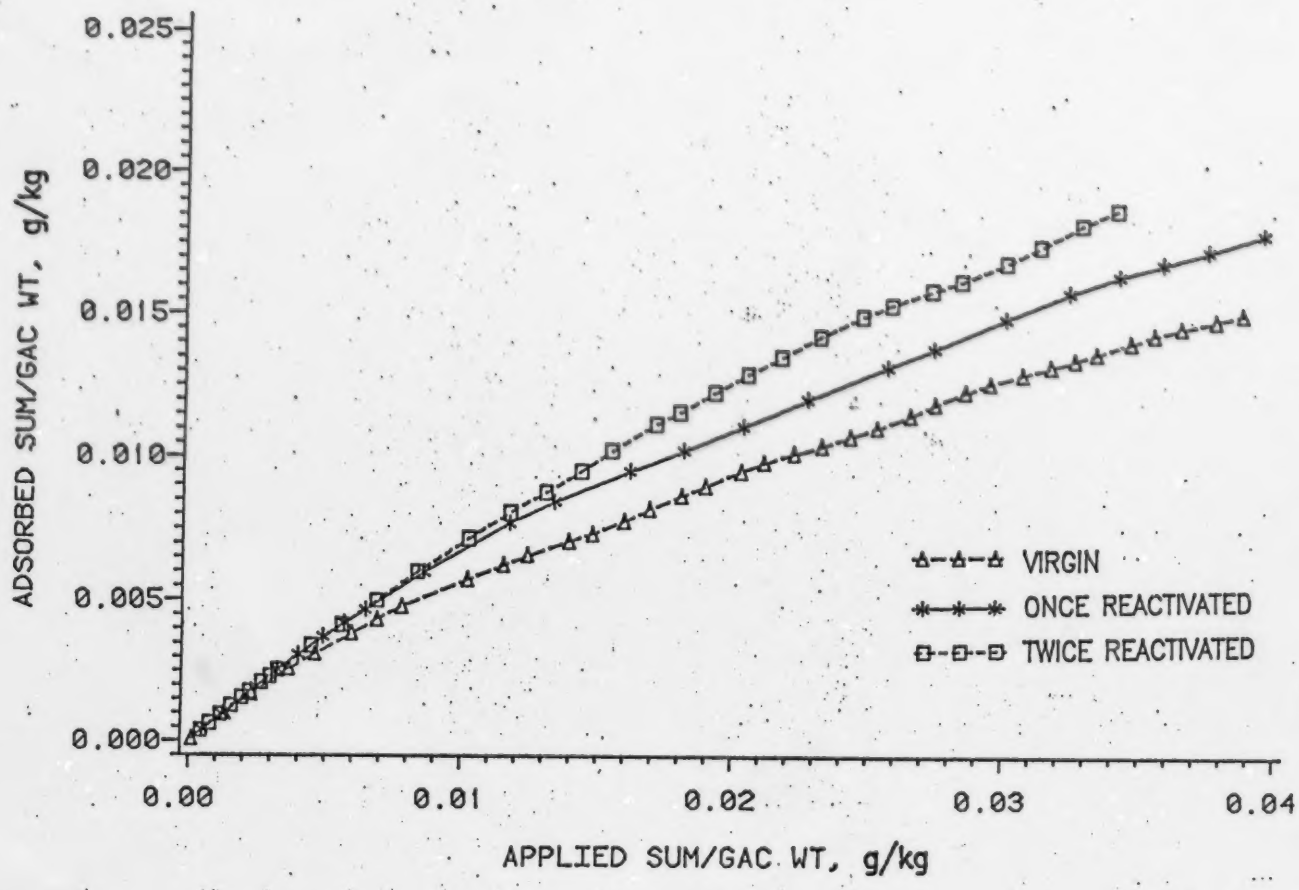

FIGURE 3. TOC LOADING ON VIRGIN AND SUBSEQUENT REACTIVATED GAC – MANCHESTER, NH.

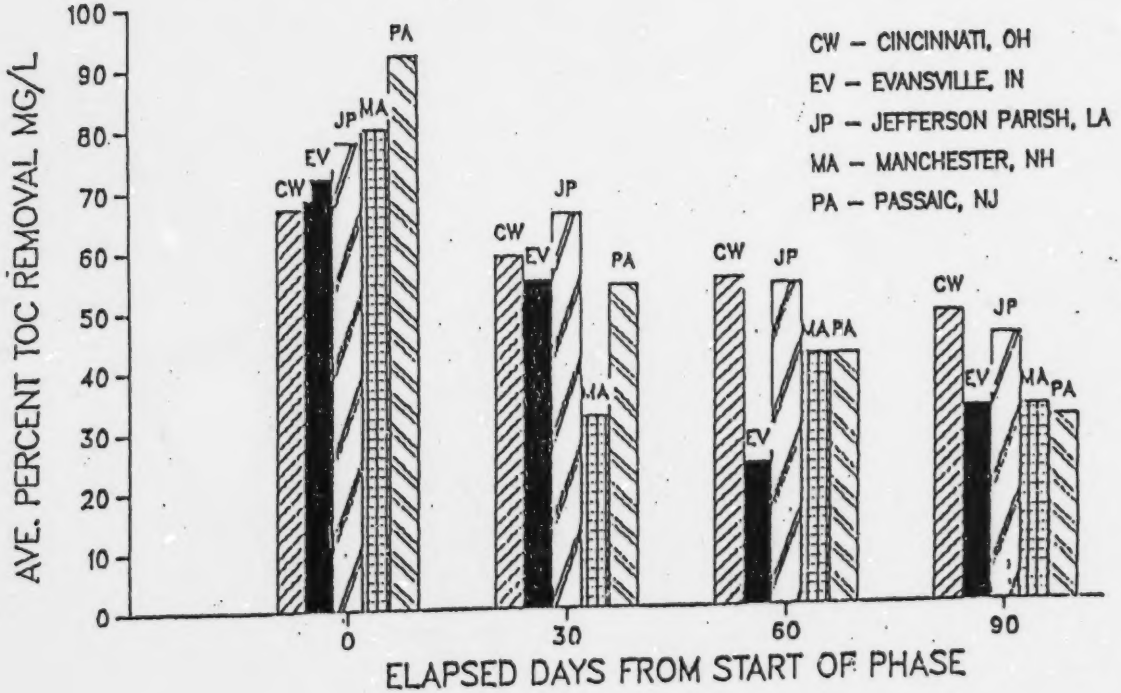

FIGURE 4. COMPARISON OF TOC REMOVAL FOR VIRGIN GAC

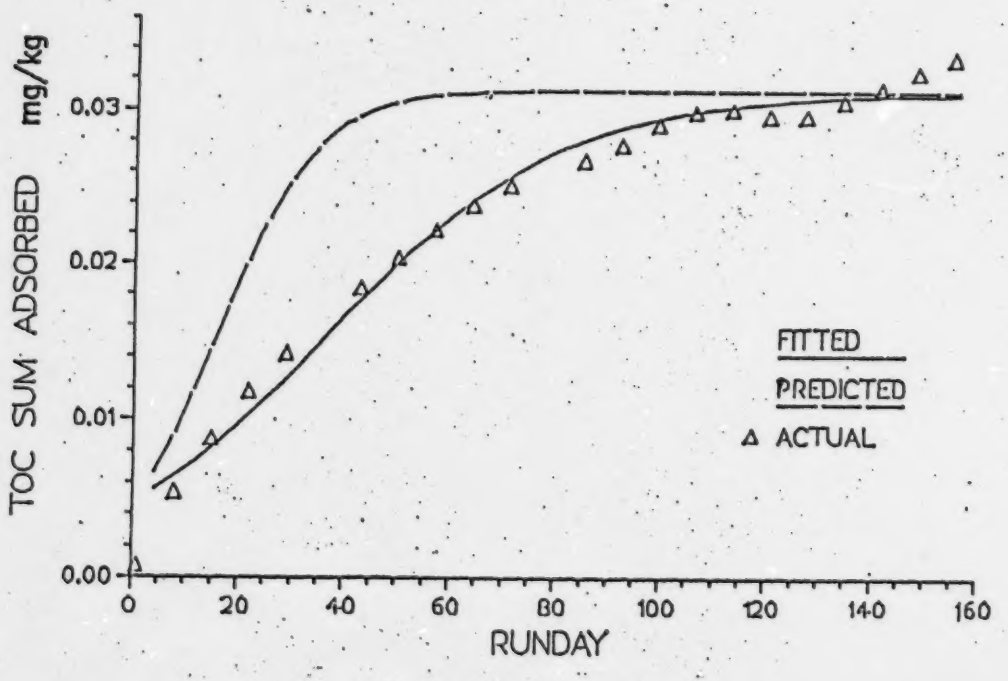

FIGURE 5. LOADING CURVES FOR VIRGIN GAC FILTER ADSORBER, CINCINNATI, OH.